COMP47750
Machine Learning with Python

2021 Assignment 2

Missing Values

Objective

The objective of this assignment is to explore how missing values can be handled in supervised machine learning. Two strategies will be explored, (1) consider missing values explicitly in the classification algorithm (Gaussian Naive Bayes) (2) use imputation methods to guess missing values.

Requirements

You may use the code from your submission for the first assignment as your starting point or you may use the sample solution for that assignment that will be provided.

1. Extend the Gaussian Naive Bayes code so that it handles missing values. Gaussian Naive Bayes can handle missing values in training by calculating conditional probabilities on the values that are present. You may choose to put a limit on the number of missing values allowed.

   Your code should also handle missing values on any test data. The easiest way to do this is to leave features with missing values out of the posterior probability calculation.

   Comment on any design decisions you make in markup.

2. Test the performance of your implementation against the scikit-learn GaussianNB using missing value imputation. Test two imputation options, one univariate and one multi-variate. To help with your evaluation two versions of the penguins datasets with missing values are provided, one with 20% missing and the other with 40%.

   You should use cross validation for testing, taking care that any scaling and imputation is handled properly within cross validation.

   Comment on the results of your evaluation.

Submission: This is an individual (not group) project. Submission is through the Brightspace page. Your submission should comprise your notebook and any datasets that you use. Clear all outputs in the notebook before saving for submission. You should use markdown cells in the notebook to report your findings and conclusions.